Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
15 July 2019

REVISED
25 September 2019

ACCEPTED FOR PUBLICATION
16 October 2019

PUBLISHED
4 February 2020

PAPER

Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials

Cas van der Oord[1,3], **Geneviève Dusson**[2,3], **Gábor Csányi**[1] and **Christoph Ortner**[2]

[1] Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
[2] Mathematics Institute, University of Warwick, Coventry CV47AL, United Kingdom
[3] These authors contributed equally to the work.

E-mail: casv2@eng.cam.ac.uk, gc121@cam.ac.uk, g.dusson@warwick.ac.uk and c.ortner@warwick.ac.uk

Keywords: materials modelling, body order expansion, data driven interatomic potentials, permutational invariant polynomials

Supplementary material for this article is available online

Abstract

We investigate the use of invariant polynomials in the construction of data-driven interatomic potentials for material systems. The 'atomic body-ordered permutation-invariant polynomials' comprise a systematic basis and are constructed to preserve the symmetry of the potential energy function with respect to rotations and permutations. In contrast to kernel based and artificial neural network models, the explicit decomposition of the total energy as a sum of atomic body-ordered terms allows to keep the dimensionality of the fit reasonably low, up to just 10 for the 5-body terms. The *explainability* of the potential is aided by this decomposition, as the low body-order components can be studied and interpreted independently. Moreover, although polynomial basis functions are thought to extrapolate poorly, we show that the low dimensionality combined with careful regularisation actually leads to better transferability than the high dimensional, kernel based Gaussian Approximation Potential.

1. Introduction

There is a long and successful history of using empirical interatomic potentials for the simulation of materials [1]. One approach is to treat such models as purely phenomenological, setting out a few key features of the true interaction to be captured (e.g. the stability ordering of certain phases), and investigate what other properties, both macroscopic or indeed microscopic, follow from these. More recently, as electronic structure calculations, particularly density functional theory (DFT) [2], has increased both in accuracy and availability, there has been a widely shared desire for potentials to match the Born–Oppenheimer potential energy surface as closely as possible [3, 4]. This change of attitude in the materials simulation community has come rather later than the analogous one in the world of organic force fields, partly due to the more systematic nature of quantum chemistry methods applicable to small molecules [5, 6].

Already a decade ago, it was clear that empirical potentials had reached their limits in terms of their ability to match the potential energy surface of DFT, essentially due to the use of simple, physically interpretable functional forms At around the same time significant developments started in which models with thousands of free parameters (so-called *high-dimensional* models) were fitted to electronic structure data. The methods are borrowed from machine learning, e.g. artificial neural networks (ANN) [7–9] and Gaussian processes (GP) [10, 11]. Although formally these models contain many degrees of freedom, they are often called *nonparametric*, because there are either good recipes for determining the best parameters that fit the data (in the case of training ANNs), or linear algebra expressions in the case of GPs. The few model parameters that are still adjusted by hand or by other ad hoc recipes are called *hyper-parameters*, e.g. the nonlinear transfer function of ANN units, or the kernel shapes in GPs. It was understood early on, similarly to the more traditional applications of machine learning, that it is advantageous to use an appropriate representation of the input data (atomic positions in this case) that captures all the known symmetries present in the problem [12]. These models achieve very high

accuracy on the training datasets and, when carefully used, on configurations that are 'near' the training, e.g. in a molecular dynamics run under similar conditions. However, the transferability of such models can still be poor. A recent attempt at assembling and fitting to a very large and diverse training dataset of elemental silicon [13], while generally staying physically sensible away from the training data, still showed up to 20% error in formation energies and migration barriers of some defects that were not in the training set. The data requirements to achieve even this level of transferability are expected to grow significantly for multicomponent systems.

While a better choice of representations and kernel functions may improve the transferability somewhat, it is conceivable that high-dimensional fits will, by their very nature, always suffer from this problem.

Similar effects can be seen in a related field, the fitting of the potential energy surfaces of molecules to high level, wave function based quantum chemistry calculations. This endeavour has a rich history [14–19], which also includes high-dimensional nonparametric fits that are very accurate for small systems (a handful of atoms), yet it is recognised that once the dimensionality reaches a few tens, the fitting task becomes extremely difficult.

At present, the only plausible way to break the curse of dimensionality is to explicitly or implicitly identify low-dimensional structures of the potential energy surface. If this can be done explicitly then the energy can be broken up into multiple low-dimensional terms, ideally ensuring that higher dimensional terms account for less variation. The challenge is to do this generally, systematically, and without sacrificing accuracy.

A time-tested and obvious way to introduce low dimensional terms is to use the *body order expansion* applied in an atom-by-atom fashion [20–25], i.e. define the total energy as a sum of one-atom, two-atom (pair), three-atom (angle) terms, and so forth. Let $\mathbf{R} \equiv \{\mathbf{r}_j\}_{j=1}^{M}$ be the positions of M atomic nuclei of the same species, representing a *configuration* of atoms (perhaps but not necessarily with periodic boundary conditions), then we write the total energy as

$$E(\mathbf{R}) = \sum_j E_1(\mathbf{r}_j) + \frac{1}{2}\sum_{i\neq j} E_2(\mathbf{r}_i, \mathbf{r}_j) + \frac{1}{3!}\sum_{i\neq j\neq k} E_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \cdots + \frac{1}{N!}\sum_{i_1\neq\ldots\neq i_N} E_N(\mathbf{r}_{i_1},\ldots,\mathbf{r}_{i_N}). \quad (1)$$

Note that if the n-body function E_n is permutation invariant then $\frac{1}{n!}\sum_{i_1\neq\ldots\neq i_n}$ may of course be rewritten as $\sum_{i_1<\ldots<i_n}$. Periodic boundary conditions are treated by taking into account the periodic images of atoms within the computational cell in (1).

We must strongly emphasize the distinction, on the one hand, between such a decomposition of the total energy into additive components, each of which only depends on few coordinates, and on the other hand the construction of *complete, many-body, invariant representations* of $\mathbf{R}$ using various combinations of two- and three-body functions, which are subsequently used in a single high-dimensional nonlinear nonparametric fit [26–28]. For example, consider the symmetry functions of Behler and Parrinello [29]. Although each element of the descriptor is itself built out of just interatomic distances (2-body) or angles (3-body), the entire descriptor vector taken as a whole is a high dimensional description of the neighbour environment of an atom, and subsequent fits are of functions in that high dimensional space. In contrast, the individual energy terms in (1) are all low dimensional, and it is to be expected that they can be fitted using much less data, since low dimensional spaces can be covered comprehensively by a relatively small number of training configurations. In particular we conjecture excellent extrapolation properties. Note that this does not require an *a priori* definition or calculation of these terms, the fitting is still to be made to total energy E and its derivatives (forces and stresses) corresponding to configurations with many atoms.

The utility of additive body order expansions has been recognised in the context of the recent machine learning based potentials. By writing the total energy as a sum of pair, triplet and a many-body terms with explicit weight factors [11, 13, 30], it is possible to prevent catastrophically erroneous predictions at small interatomic distances. Note that the failure of high dimensional fits at small interatomic distances is well known by the quantum chemists who fit small molecule potential energy surfaces [31].

However, merely writing the total energy as a body order expansion does not in itself bring the benefits of low dimensionality. The terms in (1) may be highly redundant, e.g. a general three-body potential includes all possible two-body potentials by simply not depending on one of its arguments, and this is true for all orders. A potential fitting methodology whose terms are intrinsically body-ordered was introduced as the Moment Tensor Potentials (MTPs) [32], although it would appear that no explicit use is made of the lowest dimensional terms to maximise transferability. Recently, the Atomic Cluster Expansion (ACE) was introduced and the connection between the body order expansion and the high dimensional representations used in the earlier many-body fits was also made formally explicit [33]. There, transferability is achieved by defining and calculating the low order terms in (1) explicitly using the total energy function of small clusters: the order M term is the interaction energy of the M-atom cluster in vacuum (with lower order interactions subtracted).

In both the MTP and the ACE approaches, a basis of symmetric polynomials is introduced, whose elements are body-ordered, and a linear fit in this basis is the fundamental modelling tool. The high computational cost of the many-body expansion is avoided by taking the entire many-body environment of each atom as a spatial density, and projecting it onto a rotationally invariant basis set, the components of which turn out to be

body-ordered. This density trick is the same that is used by high-dimensional descriptors, including SOAP-GAP [12], Behler–Parrinello symmetry functions [29], and the bispectrum [34, 35], but without taking advantage of the body-ordered decomposition.

Finally, it is notable that yet another route to the atomic body ordered expansion is afforded by the Generalised Pseudopotential Theory approach of Moriarty [36–38]. GPT treats perturbations of the electron density within the framework of DFT using the uniform electron gas as reference, and via a series of approximations derives individual body ordered terms formally—but also including the unit cell volume as a variable. It has had considerable success in modelling single species defect-free metals in an essentially parameter-free manner. Its extension to more complex materials is not straightforward, and calculating high body order terms (3- and 4-body terms) is rather complicated even for the simpler cases. Nevertheless GPT can be thought of as the explanation (or *justification*) of why the body order expansion is a good idea for strongly bound condensed phase materials, especially metals.

In this work, we will consider each body ordered term as an independent function to be fitted, but only whose sum is known. We will define a basis set with which we do a linear fit, so at the end, all the unknown basis coefficients may be determined in a single linear least squares problem, but separate and distinct distance based cutoffs and regularisation strategies are applied to each term. We will also forgo the vacuum-cluster definition of each term, and instead fit the basis coefficients directly to condensed phase data only. The advantage in fitting the body-order expansion to a condensed phase training set is that it can be expected to converge faster. Indeed, there is significant empirical evidence for this, such as the relative success of few-body interatomic potentials, cluster expansion for alloys [39], GPT [36–38], however, we are not aware of any rigorous results that explicitly show this in general.

For defining a basis set for the body ordered terms, we employ the theory of *permutationally invariant polynomials* (PIP) a technique based on classical invariant theory, introduced to molecular modelling for fitting the potential energy surfaces of small molecules; see e.g. [18, 23, 40] and references therein. To adapt this formalism for condensed phase covalent materials we make two modifications: (i) an explicit introduction of distance-based cutoffs into the basis functions, and (ii) each body-ordered term has its own set of PIPs, taking account of the specific symmetry group of that term. By contrast, in the original application of PIPs to small molecules, each molecule (or a set of small molecules taken as a cluster) had its potential energy surface defined and fit with the appropriate set of PIPs and each new molecule or molecular cluster required a completely new fit. In a somewhat similar vein to our ideas here, the possibility to apply PIPs to manually determined subsets (fragments) of a molecule has previously been suggested in [40] and very recently first explored in [41].

We will call our basis set 'atomic PIPs' or aPIPs, to emphasize that the body order expansion inherent to the use of PIPs is done here on an atomic rather than molecular basis. For the sake of simplicity, we limit the exposition here to elemental materials, but the formalism generalises naturally to multiple species, exactly in the same way as the original PIP formalism does for molecules.

Our overarching goal, for which the aPIP fits serve as examples, is to demonstrate for the case of materials (rather than isolated molecules) that (i) low body order potentials can reach the same high accuracies that high dimensional fits can, and (ii) polynomial fits can be used to improve generalisation properties of interatomic potentials and that sophisticated regularisation is the key to achieving this.

2. Symmetric polynomial basis

Our starting *assumption* is that (1) can be used to construct high-accuracy PESs for moderate to low body-order N, which is certainly the case empirically. Next, we require that the individual contributions E_n inherit rotation and permutation invariance (RPI) from E. Our aim is then to construct systematically improvable functional forms to represent individual E_n functions that exactly preserve these symmetries:

1. Construct coordinate systems

$$\mathbf{x}_n = \mathbf{x}_n(\mathbf{r}_1, \ldots, \mathbf{r}_n)$$

that are continuous, rotation and permutation invariant, and represent

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = \Phi_n(\mathbf{x}_n(\{\mathbf{r}_i\})).$$

2. Choose basis sets $\{B_{nj}\}$ and represent

$$\Phi_n(\mathbf{x}) = \sum_j c_{nj} B_{nj}(\mathbf{x}).$$

3. Apply a cut-off mechanism to prevent inclusion of clusters with atoms that are very far from each other that have negligible contributions to the total energy.

4. Use regularised linear least squares (i.e. ridge regression, force matching [3]) to determine the coefficients $\{c_{nj}\}_{n,j}$, using total energies, forces and stresses calculated by a first principles electronic structure approach as training data.

Several aspects of this strategy are familiar from recent machine-learning models: for example, the motivation for employing a RPI coordinate system is the same as for the use of symmetry functions descriptors for ANN potentials of Behler and Parrinello [9] as well as the SOAP descriptor and kernel of the GAP framework of Bartók and Csányi [11]. Employing a polynomial basis to obtain a systematically improvable functional form was also proposed by Braams and Bowman [40], Shapeev [42]. We will import much of the invariant theory methodology for the representation of symmetric polynomials from Braams and Bowman [40]. In the following we will show how this approach leads to a large design space and in particular describe different coordinate systems as well as the basis functions to represent the potential energy.

2.1. Distance-based potentials

Our first construction of a RPI coordinate system is achieved by closely following the ideas of Braams and Bowman [40]: we first choose rotation invariant (RI) coordinates of transformed interatomic distances,

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = E_n^{\mathrm{D}}(\{u_{ij}\}_{i<j=1}^n),$$

where u_{ij} denotes a *distance transform* such as euclidean distance itself or the common Morse or inverse distance variables,

$$u_{ij} = r_{ij}, \qquad u_{ij} = e^{-\alpha r_{ij}}, \qquad u_{ij} = r_{ij}^{-p},$$

where $\alpha,\ p > 0$. The superscript 'D' in E_n^{D} indicates that the arguments of the function are distances. Alternatively, distances and angles can be used, which we discuss below and will denote with the superscript 'DA'.

In the case of the 2-body term, u_{12} is already a RPI coordinate system. For the 3-body contribution, the permutation invariance of E_3 with respect to exchanging atom indices gives rise to full S_3 permutation invariance of E_3^{D}. Here, and throughout, S_n denotes the symmetric group over n elements. The elementary symmetric polynomials therefore give a permutation invariant coordinate system,

$$E_3(\mathbf{r}_1,\ \mathbf{r}_2,\ \mathbf{r}_3) = E_3^{\mathrm{D}}(u_{12},\ u_{13},\ u_{23}) = \Phi_3^{\mathrm{D}}(f_{3,1},\ f_{3,2},\ f_{3,3}),$$

where

$$\begin{aligned} f_{3,1} &= u_{12} + u_{13} + u_{23}, \\ f_{3,2} &= u_{12}u_{13} + u_{12}u_{23} + u_{13}u_{23}, \\ f_{3,3} &= u_{12}u_{13}u_{23}. \end{aligned} \tag{2}$$

The key property of the coordinates $f_3 = (f_{3,1}, f_{3,2}, f_{3,3})$ is that they completely define $(\mathbf{r}_1,\ \mathbf{r}_2,\ \mathbf{r}_3)$ up to rotations and permutations. The choice of such invariant coordinates is not unique. For example, we may choose $f_n = \sum_{i<j} u_{ij}^n$, $n = 1,\ 2,\ 3$ and in section 2.2 we replace distance coordinates with distance and angle coordinates.

For the n-body terms with $n \geqslant 4$, the permutation group S_n acting on $\{\mathbf{r}_i\}_{i=1}^n$ which encodes the symmetry of E_n induces a non-trivial permutation group $S_n^{\mathrm{D}} \subsetneq S_{n(n-1)/2}$ acting on $\{u_{ij}\}_{i<j=1}^n$ encoding the symmetry of E_n^{D}. That is, a permutation $\pi \in S_n$ of sites $\{\mathbf{r}_i\}$ is re-interpreted as a permutation in S_n^{D} of distances through the action

$$u_{ij} \overset{\pi}{\mapsto} u_{\pi i \pi j}$$

see figure 1 for a visualisation in the four-body case.

Employing invariant theory techniques [40, 43] we then construct *fundamental invariants* $f_n = \{f_{n,a}\}_{a=1}^{A_n}$, where $A_n > n$ is the dimensionality of the set f_n and each $f_{n,a}$ is a multi-variate polynomial in $\{u_{ij}\}$ that is invariant under S_n^{D}, such that we can rewrite E_n as

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = E_n^{\mathrm{D}}(\{u_{ij}\}) = \Phi_n^{\mathrm{D}}(\{f_{n,a}\}).$$

A subtle point is that admissible arguments $\{u_{ij}\}$ and $\{f_{n,a}\}$ belong to $(3n-6)$-dimensional submanifolds of, respectively, $\mathbb{R}^{n(n-1)/2}$ and $\mathbb{R}^{A_n}$, where $n(n-1)/2 > (3n-6)$ for $n \geqslant 5$ and $A_n > n(n-1)/2$ for $n \geqslant 4$; see table 1. For illustration, a possible choice of the fundamental invariants for a 4-body distance-based potential are given in table 2.

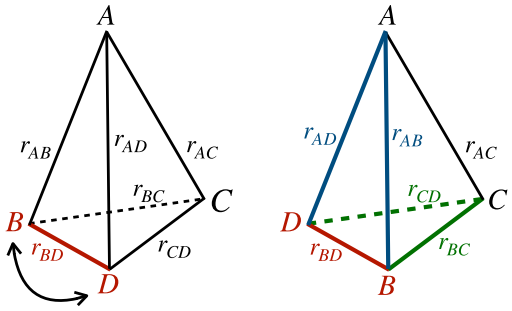


Figure 1. Permutation of the edges of a tetrahedron when two corners are permuted. In blue and green are permuted edges. Although there are six edges, the symmetry group of the edges induced by permuting pairs of atoms is a strict subset of S_6, which we denote by S_4^{D}.

Table 1. Comparison of potential minimal number of coordinates, number of RI coordinates (distances), and number of RPI coordinates (fundamental invariants). For six-body invariants MAGMA terminated without computing the fundamental invariants within several weeks of CPU time. This is related to the rapidly increasing cardinality of the symmetry group ($6! = 720$). Note that restricting the polynomial degree shortens the computations, hence allows to compute invariants for higher body-orders, as is performed in [44].

n	2	3	4	5	6
#coordinates	1	3	6	9	12
dim $\{u_{ij}\}$	1	3	6	10	15
$A_n = \dim\{f_{n,a}\}$	1	3	9	56	—

Table 2. 4B distance-based invariants.

$$
\begin{aligned}
f_{4,1} =\ & u_{12} + u_{13} + u_{14} + u_{23} + u_{24} + u_{34} \\
f_{4,2} =\ & u_{12}^2 + u_{13}^2 + u_{14}^2 + u_{23}^2 + u_{24}^2 + u_{34}^2 \\
f_{4,3} =\ & u_{12}u_{13} + u_{12}u_{14} + u_{12}u_{24} + u_{12}u_{34} + u_{13}u_{14} + u_{13}u_{23} + u_{13}u_{34} + u_{14}u_{23} + u_{14}u_{24} + u_{23}u_{24} + u_{23}u_{34} + u_{24}u_{34} \\
f_{4,4} =\ & u_{12}^3 + u_{13}^3 + u_{14}^3 + u_{23}^3 + u_{24}^3 + u_{34}^3 \\
f_{4,5} =\ & u_{12}u_{13}u_{14} + u_{12}u_{24}u_{34} + u_{13}u_{23}u_{34} + u_{14}u_{23}u_{24} \\
f_{4,6} =\ & u_{12}^4 + u_{13}^4 + u_{14}^4 + u_{23}^4 + u_{24}^4 + u_{34}^4 \\
f_{4,7} =\ & u_{12}^2u_{13} + u_{12}^2u_{14} + u_{12}^2u_{24} + u_{12}^2u_{34} + u_{12}u_{13}^2 + u_{12}u_{14}^2 + u_{12}u_{24}^2 + u_{12}u_{34}^2 + u_{13}^2u_{14} + u_{13}^2u_{23} + u_{13}^2u_{34} + u_{13}u_{14}^2 + u_{13}u_{23}^2 + \\
& u_{13}u_{34}^2 + u_{14}^2u_{23} + u_{14}^2u_{24} + u_{14}u_{23}^2 + u_{14}u_{24}^2 + u_{23}^2u_{24} + u_{23}^2u_{34} + u_{23}u_{24}^2 + u_{23}u_{34}^2 + u_{24}^2u_{34} + u_{24}u_{34}^2 \\
f_{4,8} =\ & u_{12}^2u_{13}u_{23} + u_{12}^2u_{14}u_{23} + u_{12}^2u_{23}u_{24} + u_{12}^2u_{23}u_{34} + u_{12}u_{13}^2u_{24} + u_{12}u_{13}u_{23}^2 + u_{12}u_{13}u_{24}^2 + u_{12}u_{14}^2u_{34} + u_{12}u_{14}u_{23}^2 + \\
& u_{12}u_{14}u_{34}^2 + u_{12}u_{23}^2u_{24} + u_{12}u_{23}^2u_{34} + u_{13}^2u_{14}u_{24} + u_{13}^2u_{23}u_{24} + u_{13}^2u_{24}u_{34} + u_{13}u_{14}^2u_{34} + u_{13}u_{14}u_{24}^2 + u_{13}u_{14}u_{34}^2 + \\
& u_{13}u_{23}u_{24}^2 + u_{13}u_{24}^2u_{34} + u_{14}^2u_{23}u_{34} + u_{14}u_{24}u_{34} + u_{14}u_{23}u_{34}^2 + u_{14}u_{24}u_{34}^2 \\
f_{4,9} =\ & u_{12}^2u_{13}u_{23}^2 + u_{12}^2u_{14}u_{23}^2 + u_{12}^2u_{23}^2u_{24} + u_{12}^2u_{23}^2u_{34} + u_{12}u_{13}^2u_{24}^2 + u_{12}u_{14}^2u_{34}^2 + u_{13}^2u_{14}u_{24}^2 + u_{13}^2u_{23}u_{24}^2 + u_{13}^2u_{24}^2u_{34} + u_{13}u_{14}^2u_{34}^2 + \\
& u_{14}^2u_{23}u_{34}^2 + u_{14}^2u_{24}r_{34}^2.
\end{aligned}
$$

To complete the definition of n-body distance-based potentials, we supply E_n with a cut-off mechanism, redefining them as

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = \Phi_n^{\mathrm{D}}(f_n)\prod_{i<j} f_{\mathrm{cut}}(r_{ij}), \tag{3}$$

where $f_{\mathrm{cut}}(r)$ is smooth and vanishes outside some cut-off radius r_{cut}. That is, we only account for n-body clusters for which *all* edge lengths are within the cut-off radius. The choice of cut-off mechanism is far from unique of course, and can be adapted to the systems of interest.

Table 3. 4B distance-angle-based invariants.

$f_{4,1}=$	$u_{12}+u_{13}+u_{23}$
$f_{4,2}=$	$w_{213}+w_{214}+w_{314}$
$f_{4,3}=$	$u_{12}^2+u_{13}^2+u_{23}^2$
$f_{4,4}=$	$u_{12}w_{213}+u_{13}w_{314}+u_{23}w_{214}$
$f_{4,5}=$	$w_{213}^3+w_{214}^3+w_{314}^3$
$f_{4,6}=$	$u_{12}^3+u_{13}^3+u_{23}^3+w_{213}^2w_{214}+w_{213}w_{314}^2+w_{214}^2w_{314}$
$f_{4,7}=$	$u_{12}w_{214}+u_{13}w_{213}+u_{23}w_{314}$
$f_{4,8}=$	$w_{213}^2+w_{214}^2+w_{314}^2$
$f_{4,9}=$	$u_{12}^2u_{23}+u_{12}u_{13}^2+u_{13}u_{23}^2$
$f_{4,10}=$	$u_{12}u_{13}w_{213}+u_{12}u_{23}w_{214}+u_{13}u_{23}w_{314}$
$f_{4,11}=$	$u_{12}w_{213}^2+u_{13}w_{314}^2+u_{23}w_{214}^2$
$f_{4,12}=$	$u_{12}^2w_{214}+u_{13}^2w_{213}+u_{23}^2w_{314}$
$f_{4,13}=$	$u_{12}w_{213}w_{214}+u_{13}w_{213}w_{314}+u_{23}w_{214}w_{314}$
$f_{4,14}=$	$w_{213}^2w_{314}+w_{213}w_{214}^2+w_{214}w_{314}^2$.

2.2. Distance-angle potentials

While distance-based coordinates are seemingly canonical in that they inherit the maximum symmetry, it is sometimes intuitive to employ a coordinate system that incorporates more 'physically natural' coordinates, which may lead to alternative RI and RPI coordinate systems. For example, the success of employing bond-angle coordinates in empirical force fields [22] suggests that using these coordinates may produce better fits at similar cost. This idea is further supported by the success of MTPs [42], which can be also interpreted as distance-angle potentials. Within our framework, many-body distance-angle potentials are constructed as follows.

As in section 2.1 let u_{ij} denote transformed distance coordinates. In addition, let w_{ijk} denote angle coordinates, the canonical choice being

$$w_{ijk} = \cos\theta_{ijk} = \hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ik},$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and $\hat{\mathbf{r}} = \mathbf{r}/|\mathbf{r}|$. We then write an n-body term E_n in a way that retains only partial symmetry,

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = E_n^{\mathrm{DA}}(\{u_{1j}\}_{j=2}^n, \{w_{1jk}\}_{j<k=2}^n),$$

where the superscript 'DA' indicates that the function E_n^{DA} is parametrised by distances and angles. The 2-body contribution E_2 is again a pure distance-based potential.

For body order $n \geqslant 3$, we transform again to an RPI coordinate system. For the distance-angle potentials we retain only permutation invariance with respect to the $n-1$ neighbours, which induces a different symmetry group S_{n-1}^{DA} on the coordinates $(\{u_{1j}\}_j, \{w_{1jk}\}_{jk})$. Analogously to the distance based potentials, the fundamental polynomial invariants for this permutation group S_{n-1}^{DA} yield a RPI coordinate system $f_n = f_n(\{u_{1j}\}_j, \{w_{1jk}\}_{jk})$ and thus the representation

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = \Phi_n^{\mathrm{DA}}(f_n).$$

For 3-body potentials a possible choice of fundamental invariants is

$$\begin{aligned} f_{3,1}(u_{12}, u_{13}, w_{123}) &= u_{12} + u_{13}, \\ f_{3,2}(u_{12}, u_{13}, w_{123}) &= u_{12}u_{13}, \\ f_{3,3}(u_{12}, u_{13}, w_{123}) &= w_{213}, \end{aligned} \tag{4}$$

while a possible set for four-body potentials is given in table 3.

The number of fundamental invariants is higher than for the distance based coordinate system, while the degrees of the invariants are smaller. Thus, the invariants are cheaper to compute but more basis functions will be required. This is due to the fact that we exploited less symmetry in the distance-angle potentials than in the purely distance based potentials.

Finally, we propose a natural cut-off mechanism for distance-angle potentials,

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = \Phi_n^{\mathrm{DA}}(f_n) \prod_{j=2}^{n} f_{\mathrm{cut}}(r_{1j}). \tag{5}$$

This cut-off mechanism is different from the one in the distance-based potential (3) since it only acts on the distance variables and thus the product is taken over a smaller set of r_{ij} values, leading to a summation over a

different set of n-body clusters in the total potential energy assembly. In particular, the meaning of the cutoff radius of f_{cut} in (3) and (5) is not equivalent.

2.3. Polynomial approximation

In the following, we write Φ_n to mean Φ_n^{D} or Φ_n^{DA} depending on whether we are considering distance or distance-angle coordinates. To construct computable representations of the E_n^{RPI} we will use multi-variate polynomials: the n-body functions E_n will be represented as polynomials of the invariants $f_n = \{f_{n,j}\}_{j=1}^{A_n}$, i.e.

$$E_n(\{\mathbf{r}_i\}) = \Phi_n(f_n) = P_n(f_n),$$

where P_n is a multi-variate polynomial in f_n with coefficients that are to be determined; see section 2.4. By increasing the polynomial degree of P_n we can in principle approximate arbitrary smooth, symmetric functions E_n. In particular, letting the body-order n, the cutoff r_{cut} and the polynomial degrees all tend to infinity we can represent an arbitrary smooth PES. That is, our construction is *systematically improvable*.

However, we briefly discuss a subtlety that arises for $n \geqslant 4$ when the permutation groups become non-trivial: in this case, the representation of a given symmetric polynomial in terms of the fundamental invariants is not unique. This non-uniqueness can be avoided by introducing an alternative set of invariants, the *primary invariants* $p_n = \{p_{n,a}\}_{a=1}^{n(n-1)/2}$ and *secondary invariants* $s_n = \{s_{n,b}\}_b$ both of which can be constructed from the fundamental invariants [40, 43]. In terms of p_n, s_n

$$E_n(\{\mathbf{r}_i\}_{i=1}^n) = P_n(f_n) = \sum_b s_{n,b} P_{n,b}(p_n), \tag{6}$$

where each $P_{n,b}$ is a multi-variate polynomial in $n(n-1)/2$ variables of the set p_n, and the summation ranges over all secondary invariants. Once the invariant sets p_n, s_n are specified, this decomposition of a symmetric polynomial is unique, which gives a simple way to generate all symmetric polynomials with a prescribed degree. The choice of invariants p_n, s_n remains non-unique. A 'manual' construction for $n = 4$ is proposed by Schmelzer and Murrell [45], however, for $n > 4$ this can practically only be achieved using a computer algebra system; we employ the MAGMA software package [46].

We briefly describe the primary and secondary invariants for 3- and 4-body terms For 3-body potentials, the permutation group is trivial (all of S_3), hence the $p_3 = f_3$ and $s_3 = (1)$. For 4-body potentials, the primary invariants are

$$p_4 = \{f_{4,a}\}_{a=1}^6,$$

which thus depend on the choice of coordinate system, D or DA, through the definition of f_n. The secondary invariants also depend on the choice of RI coordinates. For distance-based potentials there are six secondary invariants,

$$s_4 = (1, f_{4,7}, f_{4,8}, f_{4,9}, f_{4,7}^2, f_{4,8} f_{4,9}),$$

while for distance-angle potentials there are twelve secondary invariants,

$$s_4 = (1, f_{4,7}, f_{4,8}, \ldots, f_{4,14}, f_{4,7} f_{4,8}, f_{4,8}^2, f_{4,10}^2).$$

Note that the constant polynomial 1 is usually not considered as a secondary invariant; we include it for notational convenience.

For five-body potentials in distance based coordinates, we find 144 secondary invariants out of which 21 are irreducible. For distance-angle potentials there are 266 secondary invariants among which 44 are irreducible. Due to the complexity and large numbers of these polynomials we use the MAGMA output to *auto-generate* source-code that evaluates the invariants and their derivatives for our aPIP implementation.

2.4. Least-squares fit

All variants of the body-ordered interatomic potentials $E(\mathbf{R})$ we proposed in the foregoing sections can be expressed as a linear combination of basis functions $B_{nb\mathbf{k}}$, where n is the body-order, $\mathbf{k} \in \mathbb{N}^{n(n-1)/2}$ defines a monomial in the primary invariants and $b \in \mathbb{N}$ denotes the indices of the secondary invariant in (6).

To see this, we first specify a total polynomial degree $D_n > 0$, and recall that the invariants $p_{n,a}$ and $s_{n,b}$ are themselves polynomials in the RI coordinates and have therefore a well-defined total degree $\deg(p_{n,a})$, $\deg(s_{n,b})$. We can now write the polynomials $P_{n,b}$ from (6) as

$$P_{n,b}(\{p_{n,a}\}) = \sum_{\{k_a\}} c_{nb\mathbf{k}} \prod_{a=1}^{n(n-1)/2} (p_{n,a})^{k_a},$$

where the summation ranges over all tuples $\mathbf{k} = (k_a)_{a=1}^{n(n-1)/2}$ of non-negative integers with

$$\deg(s_{n,b}) + \sum_a k_a \deg(p_{n,a}) \leqslant D_n.$$

The coefficients $c_{nb\mathbf{k}}$ are the unknowns to be determined in the least squares fit. Thus we see that for each body-order n, each secondary invariant indexed by b and for each tuple $\mathbf{k}$ specifying a monomial within the prescribed degree D_n we obtain a corresponding basis function for the total potential energy

$$B_{nb\mathbf{k}}(\{\mathbf{r}_i\}_{i=1}^M) = \sum_{\substack{1<i_1<i_2 \\ <\ldots<i_n\leqslant M}} F_{\text{cut}}(\{\mathbf{r}_{i_l}\}_{l=1,\ldots,n}) \times \left[s_{n,b} \prod_{a=1}^{n(n-1)/2} (p_{n,a})^{k_a} \right], \tag{7}$$

where $s_{n,b}$, $p_{n,a}$ are evaluated at $\{\mathbf{r}_{i_l}\}_{l=1,\ldots,n}$ and the definition of the cut-off function $F_{\text{cut}}(\{\mathbf{r}_{i_l}\}_{l=1,\ldots,n})$ depends on the choice of variables and is defined in (3) for the distance-based case and (5) for the distance-angle case. In the summation, only clusters respecting the cut-off condition $F_{\text{cut}}(\{\mathbf{r}_{i_l}\}_{l=1,\ldots,n}) > 0$ taken into account to ensure linear scaling cost.

It remains to determine the coefficients $c_{nb\mathbf{k}}$ in the linear expansion

$$E(\{\mathbf{r}_i\}_{i=1}^M) = \sum_{n,b,\mathbf{k}} c_{nb\mathbf{k}} B_{nb\mathbf{k}}(\{\mathbf{r}_i\}_{i=1}^M), \tag{8}$$

achieved via solving a linear least squares problem.

For each atomic configuration $\mathbf{R}$ in a training set $\mathcal{R}$, the corresponding energy $\mathcal{E}_R$, forces $\mathcal{F}_\mathbf{R}$ and possibly virials $\mathcal{V}_\mathbf{R}$ are given. The minimized functional is of the form

$$J = \sum_{\mathbf{R}\in\mathcal{R}} (W_E^2 |E(\mathbf{R}) - \mathcal{E}_\mathbf{R}|^2 + W_F^2 |F(\mathbf{R}) - \mathcal{F}_\mathbf{R}|^2 + W_V^2 |V(\mathbf{R}) - \mathcal{V}_\mathbf{R}|^2), \tag{9}$$

where W_E, W_F, W_V are weights that may depend on the configurations $\mathbf{R}$, and $F(\mathbf{R})$ and $V(\mathbf{R})$ are, respectively, forces and virials computed from the energy functional $E(\mathbf{R})$.

In summary, since J is quadratic in the unknown polynomial coefficients $\mathbf{c} = \{c_{nb\mathbf{k}}\}$, its minimisation is a standard linear least-squares problem

$$\min_{\mathbf{c}} \|A\mathbf{c} - Y\|_2^2, \tag{10}$$

which we solve using a QR factorisation. The size of the system matrix A is $N_{\text{obs}} \times N_{\text{basis}}$ where N_{basis} is the number of basis functions while N_{obs} is the number of observations (energies, forces, virials, and regularisation, if any, see below). In our examples N_{obs} may be in the range of hundreds of thousands, however, N_{basis} remains low; on the order of hundreds to a few thousands. In this case, the QR factorisation of the matrix A is computationally cheap ($O(N_{\text{obs}} \times N_{\text{basis}}^2)$ operations) and numerical stable. We postpone discussion of regularisation mechanisms to the next section, some of which will show up as a regularisation functional added to J, which will thus remain a quadratic functional in $\mathbf{c}$.

2.5. Systematic convergence

The prospective accuracy of an interatomic potential is directly related to its functional form, in our case the choice of basis functions to represent the PES. The family of potentials we constructed in the previous sections are systematically improvable: by increasing the body-order, cutoff radius and polynomial degree they are *in principle* capable of representing an arbitrary many-body PES to within arbitrary accuracy. To support this claim, we begin by studying the convergence of the root mean square error (RMSE) on two previously published training sets for tungsten and silicon.

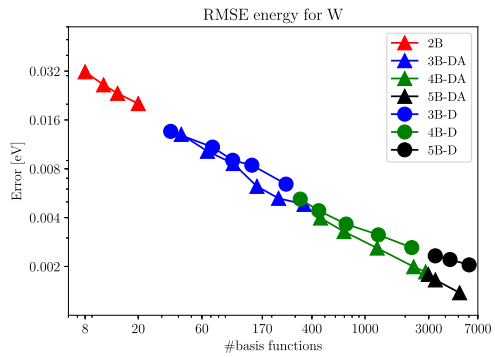
We measure the convergence of the RMSE against two key features: Firstly, the number of basis functions used to construct the potential gives a crude measure of the cost of the training. Secondly, we compare the accuracy of the fit against the evaluation time of the forces, that is, the cost of one molecular dynamics step.

For both training sets, we demonstrate the convergence of the potential for both the distance-based and the distance-angle descriptors. For all potentials, the distance transform used is a polynomial transform, that is $u_{ij} = (r_{\text{nn}}/r_{ij})^p$ where r_{nn} is an estimate for the nearest-neighbour distance ($r_{\text{nn}} = 2.74$ Å for W and $r_{\text{nn}} = 2.35$ Å for Si) and p may vary with the body-order. The cutoff function is given by

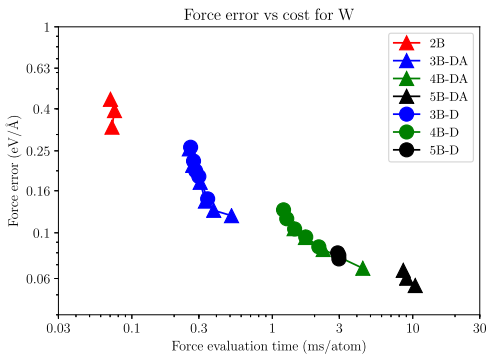
$$f_{\text{cut}}(r) = \begin{cases} [(r/r_{\text{cut}})^2 - 1]^2, & 0 \leqslant r < r_{\text{cut}}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where r_{cut} is a cut-off radius that may again vary with the body-order. The parameters for the individual potentials and the least squares regression weights are given in the supplementary material[4].

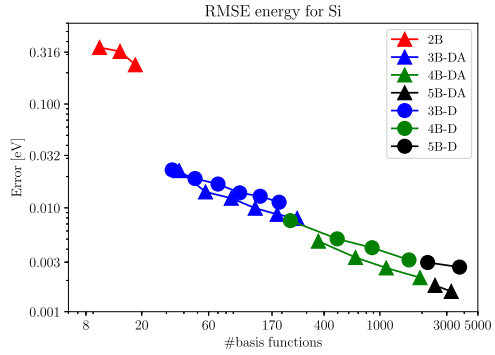
[4] See supplementary material available online at link stacks.iop.org/MLST/1/015004/mmedia for a description of the 5-body invariants, convergence tables for tungsten and silicon, and details on the titanium database generation and parameters.



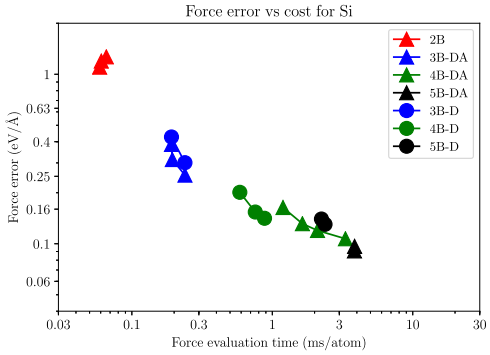
(a) Convergence of the RMS error on the energy for a tungsten database with distance-angle descriptors (triangles) and distance-based descriptors (circles).



(b) Timing for the force evaluation per atom for a tungsten database with distance-angle descriptors (triangles) and distance-based descriptors (circles).



(c) Convergence of the RMS error on the energy for a silicon database with distance-angle descriptors (triangles) and distance-based descriptors (circles).



(d) Timing for the force evaluation per atom for a Silicon database with distance-angle descriptors (triangles) and distance-based descriptors (circles).

Figure 2. Systematic convergence on a tungsten (top) and a silicon (bottom) database. Note that the SOAP-GAP model takes about 30 ms/atom to evaluate on a range of current CPUs, and MTP is 1 ms or faster depending on the number of basis functions used.

To choose the functional form of the distance transforms and cutoff function we first performed low-accuracy fits that showed that the fit accuracy varies little across different choices of cut-off function and distance transform; see the supplement (see footnote 4). However, we will find that distance-angle potentials achieve a higher accuracy at comparable computational cost than distance-based potentials, both on the W and Si training sets.

2.5.1. Results for tungsten

We now present convergence results for a tungsten training set used for a previously published SOAP-GAP model [47], generated with CASTEP [48], that consists of 9693 configurations including primitive unit cells, surfaces, γ-surfaces, vacancies and dislocation quadrupoles. Every configuration provides one total energy and $3N_{at}$ force components where N_{at} is the number of atoms per configuration. Some configurations also provide six virial components. The resulting total number of scalars used for the fit was 497271.

For both distance-based and distance-angle potentials we observe in figure 2(a) the systematic decrease of the RMSE as the body-order and the polynomial degrees are increased. Extended convergence tables are presented in the supplementary material (see footnote 4).

In this test, distance-angle potentials perform slightly better than distance-based potentials, particularly in the high accuracy regime. Indeed, the distance-based potentials with 5-body reach an energy RMSE of 2.05 meV with 6023 basis functions and 2.98 ms force evaluation time per atom, while the distance-angle potentials for 4-body reach an energy RMSE of 1.85 meV with 2842 basis functions and 4.41 ms force evaluation time per atom. Thus, both the errors and computational costs are comparable. The 5-body distance-angle potentials reach an energy RMSE of 1.38 meV with 5113 basis functions and 10.4 ms force evaluation time per atom.

2.5.2. Results for silicon

We now demonstrate the convergence of the potential on a previously published silicon training set [13] which contains 2475 diverse configurations. We restrict the published database to train only on the following subset of configurations: diamond cubic, amorphous, β-tin, vacancies, sp2, and low index surfaces. The total amount of scalars included in the fit (total energies, force/virial components) for this subset of the silicon database is 323414. Although there are fewer configurations overall than in the tungsten database, there are two distinct solid phases, and the amorphous phase which is particularly challenging to fit. On the one hand, excluding certain parts of the published database allows us to explore extrapolation. On the other hand, efficiently and accurately fitting to the complete training set including a large variety of high coordination phases will likely require a more flexible functional form, which we will revisit in future work.

The full convergence tables are presented in the supplementary material (see footnote 4).

As for tungsten, the choice of descriptors gives similar accuracy and evaluation times for silicon, but distance-angle potentials now reach significantly lower errors for large basis sets. Moreover, the convergence plots presented on figures 2(c) and (c) show a systematic convergence of the energy error for distance-angle and distance-based descriptors. More precisely, the accuracy reaches the value of 2.13 meV accuracy for a distance-angle 4-body potential composed of 1933 basis functions with a force evaluation time of 3.32 ms per atom, and for a distance-based 5-body potential, the energy error reaches 2.49 meV composed of 5396 basis functions with a force evaluation time of 2.51 ms per atom. Finally, the 5-body distance-angle potentials reach an energy error of 1.47 meV with 3759 basis functions and a force evaluation time of 3.91 ms per atom.

3. Regularisation

3.1. Regularisation techniques

In the least-squares method, regularisation is primarily seen as a procedure to improve conditioning on ill-conditioned or even ill-posed problems. By contrast, in the Gaussian process framework, it can be interpreted as imposing 'prior' information about the potential energy surface, in particular its regularity. Robust heuristics for choosing the strength of the regularisation were crucial for the success of the GAP scheme for materials, where the regulariser was chosen to be consistent with the estimated convergence error in the input data e.g. with respect to k-point sampling [49].

In the following we seek to apply a similar perspective in the standard least squares framework. We will show how the low-dimensional functional forms obtained in our definition of aPIPs in section 2 allow us to incorporate physically motivated 'prior' information or requirements that are not present in the database.

For example, consider the unregularised pair potential fit to the W database displayed in figure 3(c), which is obtained by fitting a degree 16 polynomial with distance transform $u_{ij} = (2.74\text{Å}/r_{ij})^2$ and cutoff radius $r_{\text{cut}} = 8.5$ Å; see section 3.2 for more details. While it gives low RMSE on our dataset, it is a nonsensical pair potential that is unsuitable for materials modelling work.

A typical approach to detect overfitting and validate the generalisation capabilities of the fitted potential is to first separate the data into a training set and a test set, then to perform the regression using only the training set, and finally compute the errors separately on the training and on the test set. A transferable fit should have comparable training and test errors.

While such a procedure helps to prevent overfitting *near* the training set, we find that it gives very limited information about the ability of the potential to generalise more broadly. While training and test errors are comparable for a proportion of training configurations of 0.6 and higher (figure 3(a)), all potentials exhibit an oscillatory shape (figure 3(c)) which is pathological and clearly does not allow for extrapolation. In addition, the pair potential minimum is far from the nearest-neighbour distance. Therefore, the generalisation tests presented in section 4 are performed directly on physical properties and not using a training/test splits.

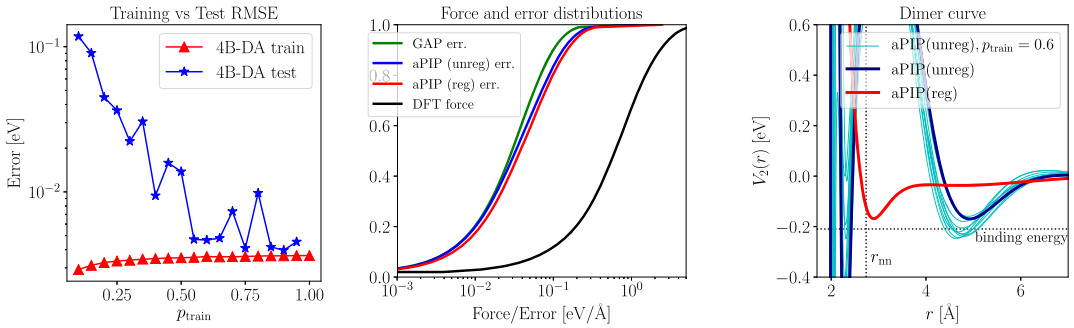
We now introduce a range of tools that enable us to produce regularised aPIP fits that retain RMSE accuracy close to the unregularised aPIPs, but become highly transferrable potentials that 'extrapolate well' and in particular have no regions of 'holes' as those described in [31].

3.1.1. Tikhonov regularisation

First, we recall some background on regularisation. In the context of linear least squares, the problem (10) is replaced with the regularised least squares problem

$$\min_{\mathbf{c}} \|A\mathbf{c} - Y\|_2^2 + \|\Gamma\mathbf{c}\|_2^2, \tag{12}$$

where Γ is called the Tikhonov matrix. The form $\|\Gamma\mathbf{c}\|_2^2$ may be used to represent any positive quadratic functional acting on the aPIP potential energy surface E given by (8). The most common choice is $\Gamma = \alpha I$



(a) Training and test errors for a series of unregularised 4B-DA potentials for tungsten. Here, p_{train} is the proportion of training configurations, chosen at random.

(b) Cumulative force error distribution on the entire tungsten training set. The aPIP potentials here correspond to $p_{\text{train}} = 1$ of the other panels, and the SOAP-GAP model is the one originally published with the data set.

(c) Two-body part of the unregularised aPIP potentials with $p_{\text{train}} = 0.6$ (multiple curves correspond to different subsets) and $p_{\text{train}} = 1.0$, as well as the regularised potential; see § III B for the details.

Figure 3. Comparison of training and test errors, and the corresponding two-body potentials.

(L^2-regularisation) where the unknown parameter α may be obtained through ad hoc procedures, or related to the uncertainty of the data via the Bayesian interpretation of the least squares problem.

Such regularisation techniques are, for example, employed to render ill-posed problems well posed, or improve the conditioning of severely ill-conditioned problems. In our context, polynomial basis functions generally lead to ill-conditioning, which is exacerbated by the fact that the space of n-body functions contains m-body functions with $m < n$, leading to a near-degeneracy that is only partly alleviated by using of a different cut-off and distance transform at each body-order.

To solve the regularised least squares problem we re-interpret it as a standard least squares problem through the equivalent formulation

$$\min_{\mathbf{c}} \left\| \begin{bmatrix} A \\ \Gamma \end{bmatrix} \mathbf{c} - \begin{bmatrix} Y \\ 0 \end{bmatrix} \right\|_2^2,$$

which is then solved using the QR factorisation.

3.1.2. Rank-revealing QR factorisation

L^2-regularisation can be effectively replaced by the rank-revealing QR factorisation (rr-QR) [50], a decomposition which reveals the near-degeneracy of the matrix A. The factorisation reads

$$AP = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix},$$

where P is a permutation matrix, Q is an orthogonal matrix, R_{11} and R_{22} are upper triangular matrices and, importantly, a given norm of the matrix R_{22} is below some prescribed tolerance.

Truncating R_{22} in the resolution of the least-square system can be seen as a regularisation as it removes the small modes in the matrix A. To demonstrate this, let us compare rr-QR and L^2-regularisation on a simple example, where A is the nearly rank-deficient matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad \varepsilon \text{ small},$$

and the observations are $Y = (y_1, y_2)^T$. The solution of the unregularised least squares problem is $c_1 = y_1$, $c_2 = y_2/\varepsilon$. The rr-QR algorithm with a parameter α will instead compute

$$c_1 = y_1, \quad c_2 = \varepsilon^{-1} \delta_{\varepsilon > \alpha} y_2,$$

while the least squares solution with L^2-regularisation with parameter α is given by

$$c_1 = \frac{1}{1 + \alpha^2} y_1, \quad c_2 = \frac{\varepsilon}{\varepsilon^2 + \alpha^2} y_2.$$

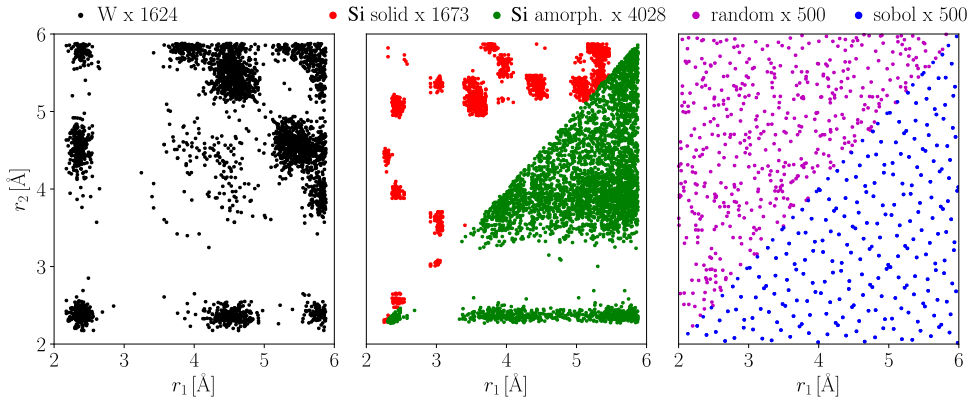


Figure 4. Distribution of 3-body clusters in the W and Si training sets versus artificially generated (random, Sobol) clusters. The clusters are represented by distance-angle variables (r_1, r_2, w_{12}) with the angle variable $w_{12} = \hat{R}_1 \cdot \hat{R}_2$ restricted to $-0.3 \leqslant w_{12} \leqslant -0.2$. The picture is similar for other angle-intervals.

The two solutions are asymptotically equivalent and tend to the unregularised solution as $\alpha \to 0$. In particular, the first coefficient is exactly right with the rr-QR factorisation but not with the Tikhonov regularisation.

3.1.3. Integral functionals

We now introduce a class of regularisers that are made possible by the fact that we decompose the PES into relatively low-dimensional components, the body-orders. A special case, discussed in the next section will be a critical ingredient in our fitting procedure.

Consider a PES given by a body-order expansion (1), and let us assume that we write E_n as a distance-based or distance-angle potential, i.e. $E_n(\{\mathbf{r}_i\}_i) = V_n(\mathbf{u}_n(\{\mathbf{r}_i\}))$. Then we consider a regularisation functional of the form,

$$\|\Gamma_n \mathbf{c}\|_2^2 = \int w(\mathbf{u}) |L[V_n(\mathbf{u})]|^2 \, d\mathbf{u}, \tag{13}$$

where L is a linear differential operator, w an integration weight and integration is taken over the domain of definition of V_n, i.e. all admissible tuples $\mathbf{u}$ that can be written as $\mathbf{u} = u_n(\{\mathbf{r}_i\})$. The right-hand side can be written in the form of a Tikhonov functional since V_n depends linearly on a subset of coefficients $\mathbf{c}$.

To approximately evaluate this integral we choose integration points $\{\mathbf{u}_j\}_{j=1}^J \subset \mathbb{R}^d$ that are distributed according to the measure $w(\mathbf{u})\,\mathrm{d}\mathbf{u}$ and replace the integral functional (13) with its discretised variant

$$\|\Gamma_n \mathbf{c}\|_2^2 = \frac{1}{J} \sum_{j=1}^{J} |L[V_n(\mathbf{u}_j)]|^2. \tag{14}$$

A canonical choice for $\{\mathbf{u}_j\}_{j=1}^J$ are low-descrepancy sequences; we simply use the classical Sobol sequence. This is effective in low and moderate dimensions where Sobol sequences 'fill space' with few (O(1000) to O(100 000)) points [51]. In principle one could also use random number sequences instead.

In figure 4 we show how solid configurations are highly concentrated in the space of n-body clusters. Amorphous configurations 'fill space' much better (liquid even more so), and can to a certain degree be seen as a 'natural' regulariser, however they still concentrate in parts of configuration space. By contrast, random or Sobol sequences provide close to uniform distributions of datapoints at which to apply the regularisation, or alternatively their concentration can be easily tuned by adjusting the upper and lower bounds or through applying a distance transformation.

3.1.4. Laplace smoother

A wide variety of choices for the differential operator L in (14) are possible. In the present work we will only consider the Laplace operator, $L = \gamma\Delta \equiv \gamma\nabla^2$, i.e. (14) becomes

$$J_n^{\Delta} = \frac{\gamma_n}{J} \sum_{j=1}^{J} |\Delta[V_n(\mathbf{u}_j)]|^2, \tag{15}$$

where γ_n is an adjustable regularisation parameter. Although it is in principle possible to implement second derivatives of V_n, we have chosen instead to approximate ΔV_n with a finite-difference,

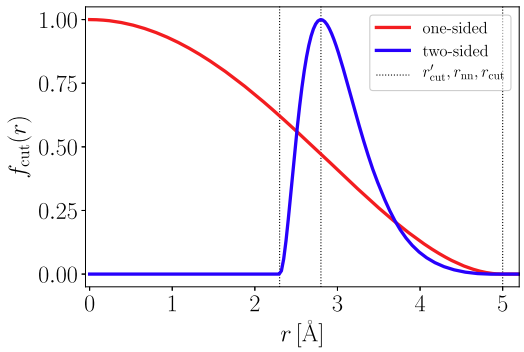


Figure 5. A one-sided cutoff function (11) for the two-body function V_2 and a two-sided cutoff function (16) for n-body functions V_n, $n \geqslant 3$. The illustrative parameters are $r'_{\text{cut}} = 2.3$Å, $r_{\text{nn}} = 2.8$Å, $r_{\text{cut}} = 5.0$Å.

$$\Delta_h V_n(\mathbf{u}) = h^{-2} \sum_{\alpha=1}^{d} (V_n(\mathbf{u} + h\mathbf{e}_\alpha) - 2V_n(\mathbf{u}) + V_n(\mathbf{u} - h\mathbf{e}_\alpha)).$$

Regularising the least squares fit with the functional J_n^{Δ} promotes that the *curvature* ΔV_n is moderate, which is a gentle requirement of smoothness. By adjusting the parameter γ_n, smoothness can be traded against accuracy of fit.

3.1.5. Two-sided cutoffs and repulsive core

For distances well below the nearest-neighbour distance regularisation is particularly crucial as demonstrated in [31] and in figure 3(c). While we could apply the Laplace regulariser in this region to control oscillations of the polynomials and thus prevent 'holes' in the PES, this would inhibit the ability of the polynomials to produce an accurate fit in regions of interest. Instead, we chose to (1) apply an inner cutoff to all V_n, $n \geqslant 3$; and (2) replace the global two-body polynomial V_2 with a spline that guarantees repulsion at short interatomic distances. In detail we apply these ideas as follows:

(a) Two-sided cutoff: Typically, the cutoff function f_{cut} appearing in (3) and (5) are positive on an interval $[0, r_{\text{cut}})$ and vanish on $[r_{\text{cut}}, \infty)$. For example, we often use the spline defined in (11).

In order to prevent oscillation and blow-up of the n-body functions V_n, $n \geqslant 3$ we require that $f_{\text{cut}} = 0$ on both $[0, r'_{\text{cut}}]$ and $[r_{\text{cut}}, \infty)$. A specific choice that we used in our tests is

$$f_{\text{cut}}(r) = \begin{cases} C(\xi^2 - 1)^2, & r_{\text{cut}}{}' < r < r_{\text{cut}}, \\ 0, & \text{otherwise}, \end{cases}$$
$$\xi = \exp(\lambda(r/r_{\text{nn}} - 1)) - 1, \tag{16}$$

where r_{nn} is an estimate for the ground state nearest neighbour distance in the material under consideration, λ is chosen such that the resulting f_{cut} has its unique local maximum at r_{nn} and C such that $f_{\text{cut}}(r_{\text{nn}}) = 1$. See figure 5 to visualise this construction. We emphasize, however, that there are many reasonable alternatives to implement this.

(b) Repulsive two-body: We initially perform the regularised least-squares fit with a global polynomial representation of V_2 as described in section 2. We then choose a spline point $r_{\text{S}} < r_{\text{nn}}$, sufficiently small so that modifying $V_2(r)$ for $r < r_{\text{S}}$, ideally chosen small enough so that the RMSEs are not significantly affected. This point must furthermore be chosen so that $V_2{}'(r_{\text{S}}) < 0$. We then define a new two-body potential

$$\tilde{V}_2(r) := \begin{cases} V_2(r), & r \geqslant r_{\text{S}}, \\ V_{\text{rep}}(r), & r < r_{\text{S}}, \end{cases}$$
$$V_{\text{rep}}(r) = e_\infty + \beta r^{-1} e^{-\alpha r},$$

where $e_\infty < V_2(r_{\text{S}})$ is a tuning parameter that can be used to adjust the steepness of the potential, while α, β are chosen such that $\tilde{V}_2$ is continuous and continuously differentiable at r_{S}. The form of the repulsive potential V_{rep} is arbitrary, and in applications where it is important to accurately describe interactions between atoms at very close distances it should be chosen as or similar to the universal ZBL function [52]. The repulsive core constructions for the regularised W and Si fits, described in detail in section 3.2, are visualised in figure 6.

In practice, the inner cutoff and splining mechanisms interact mildly with the regularised least squares regression, and we did not find it particularly difficult to find suitable parameter choices.

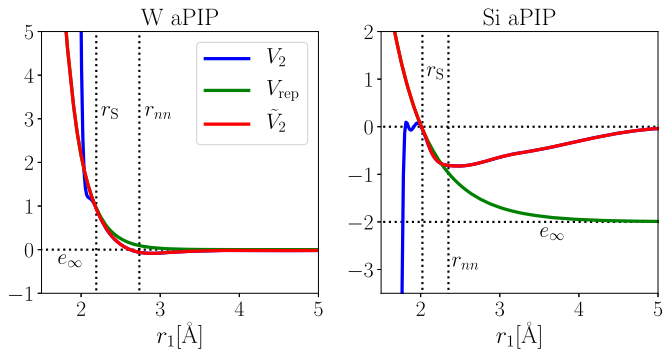


Figure 6. Spline two-body potentials with repulsive core; see section 3.1.5(b).

Table 4. RMSE accuracy on W training set, comparing SOAP-GAP [47] against regularised and unregularised aPIP.

	Energy (meV)			Forces (meV Å^{-1})			Virials (meV)		
Config type	GAP	aPIP(unreg)	aPIP(reg)	GAP	aPIP(unreg)	aPIP(reg)	GAP	aPIP(unreg)	aPIP(reg)
Unit cells	0.07	0.24	0.27	0.00	0.00	0.00	3.47	3.80	4.17
Bulk MD	0.60	0.47	0.45	27.8	21.4	26.5			
Vacancy	0.50	0.26	0.67	29.4	25.3	29.5			
Dislocation	1.86	0.98	1.02	38.3	33.0	35.7			
Surface	0.45	0.43	0.88	50.9	42.2	70.8			
γ-surface	1.66	2.92	4.09	69.0	99.4	118.0			
γ-s. vacancy	1.26	1.70	2.71	79.3	98.7	109.3			

3.1.6. Sequential fits

A source of ill-conditioning in the least squares system (10) is due to the fact that any m-body function V_m can *nearly* be represented as an n-body function V_n with $n > m$. Thus, a final mechanism that we employ is to fit different n-body terms independently from one another. For example, we may first fit a two-body V_2, followed by the modification described in section 3.1.5(2). Then, in a separate step we fit V_3, V_4, and possibly V_5 after subtracting the values for V_2 from the observations vector. At present we perform this procedure in a purely ad hoc fashion.

3.2. Accuracy of regularised aPIPs

In this section, we investigate the effect of our regularisation procedures on the fit accuracy for the W and Si training sets described in section 2.5, as well as a small selection of material properties. While the Si fits will be performed with a 5-body potential, for the W potential we restrict the body-order to four since this already achieves satisfactory accuracy on the W training set.

The majority of hyperparameters for the Si and W fits are identical and can be summarized as follows: all potentials are distance-angle potentials with the same distance transforms u_{ij} as in the RMSE convergence tests. For the unregularised aPIPs the cutoff function is (11) while for the regularised aPIPs we use the one-sided cutoff (11) only for the two-body potential but a two-sided cutoff (16) for all V_n, $n \geqslant 2$. The least squares functionals are weighted differently from the RMSE convergence tests where we were targeting *total RMSEs*. For the regularisation and extrapolation tests we chose the weights to aim for accuracy on subsets comparable to the previously published SOAP-GAP models. The regularised fits employ the complete range of tools introduced in section 3.1. The specific details of the aPIP potential parameters, fitting parameters and regularisation parameters, and sequential fitting procedure, are given in the supplement.

The resulting unregularised and regularised potentials will, respectively, be denoted by `aPIP(unreg)` and `aPIP(reg)`.

3.2.1. Results: tungsten

We compare the aPIP RMSE for energies, forces and virials per configuration type in the tungsten training set against the original SOAP-GAP model published with the data set [47]. The results can be found in table 4. The purpose of this test is to confirm that the regularisation only slightly reduces the RMSE accuracy per atom compared to the unregularised aPIP(unreg).

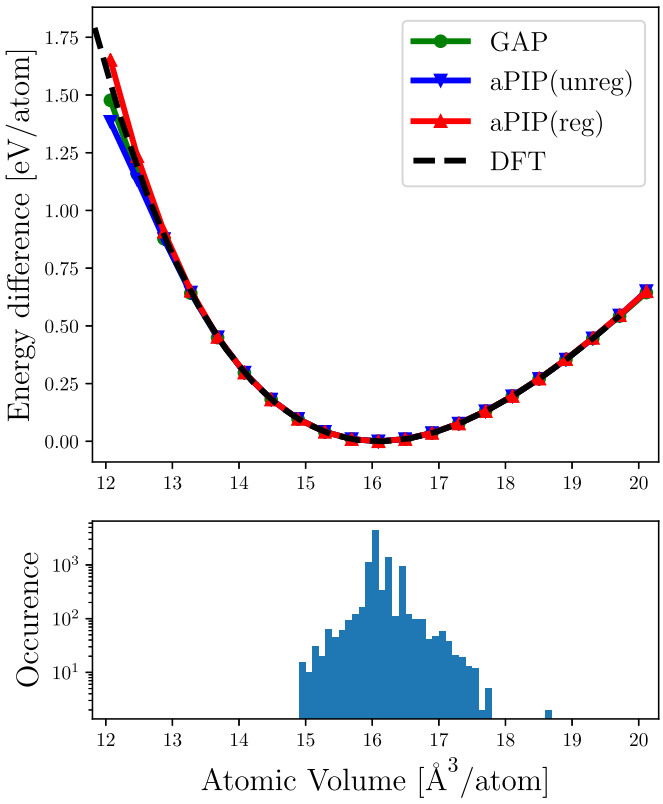


Figure 7. Tungsten bcc energy versus volume comparing SOAP-GAP [47] against regularised and unregularised aPIP. The lower panel shows the histogram of volumes in the data set.

Table 5. RMSE accuracy on Si training set, comparing SOAP-GAP [13] against regularised and unregularised aPIP.

	Energy (meV)			Forces (meV Å^{-1})			Virials (meV)		
Config type	GAP	aPIP(unreg)	aPIP(reg)	GAP	aPIP(unreg)	aPIP(reg)	GAP	aPIP(unreg)	aPIP(reg)
dia	0.65	0.33	0.55	15.4	13.9	21.8	14.59	6.09	9.96
amorph	0.56	2.08	4.49	102.7	138.6	172.0	60.9	15.85	31.32
bt	0.72	0.27	0.45	17.2	18.2	31.2	26.47	12.94	22.15
vacancy	0.54	0.36	0.71	48.3	46.2	62.7	9.82	7.52	11.89
sp2	0.48	0.70	1.82	48.6	45.0	79.2			
surface110	0.20	0.48	2.80	161.2	159.0	237.0			
surface111	0.22	0.31	0.99	156.9	155.8	233.7			
surface001	0.19	0.50	1.39	140.4	136.9	195.7			

Apart from monitoring the RMSE we also benchmark the fitted potentials by comparing their predictions for various material properties: energy–volume curves of the aPIP models as well as the SOAP-GAP model are compared against DFT in figure 7. These results are shown to be in excellent agreement for all the fitted potentials. A second test is to calculate the elastic constants B, C_{11}, C_{12}, C_{44} for a bcc tungsten structure. The aPIP (unreg), aPIP(reg) and SOAP-GAP all achieve to predict the elastic constants within 1%.

3.2.2. Results: silicon

The RMSE accuracy per configuration type of unregularised aPIP(unreg) and regularised aPIP(reg) are compared against the SOAP-GAP fit [13] in table 5. Again the regularised aPIP(reg) is shown to decrease in RMSE accuracy compared to the unregularised aPIP(unreg), and by larger factors than in the case of tungsten.

The fitted potentials were again compared for a range of different material properties. The energy versus volume curve for silicon is shown in figure 8 comparing the fitted potentials to the DFT reference, with excellent agreement for each. The elastic constants were calculated as well as the surface and vacancy formation energies and are presented in table 6. The unregularised aPIP(unreg) is shown to have larger errors on the elastic constants compared to the aPIP(reg) and most notably failed the vacancy energy test. That is, during the

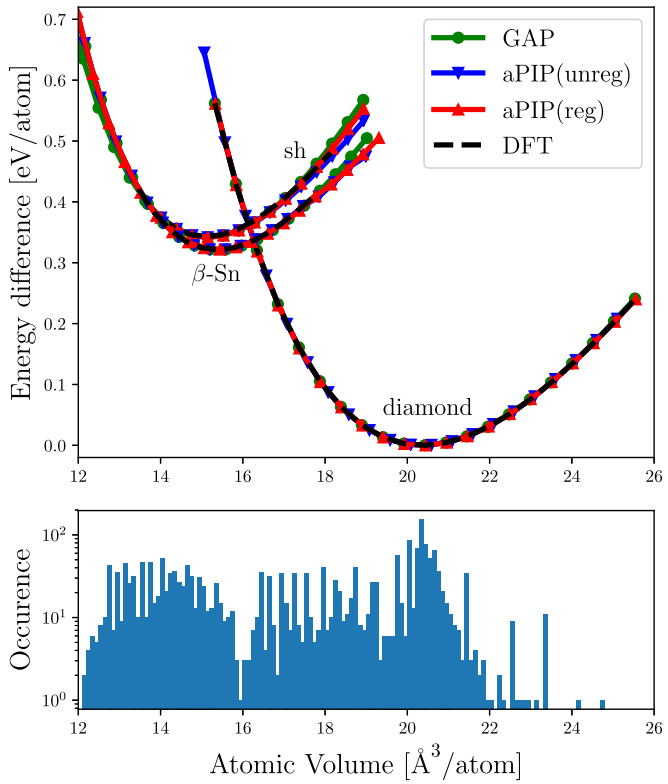


Figure 8. Energy volume curve for diamond and β-Sn silicon comparing the regularised and unregularised aPIPs, and SOAP-GAP potentials against DFT. The lower panel shows the histogram of the volumes in the data set.

Table 6. Relative error on a range of different properties for Si, comparing the SOAP-GAP, aPIP and aPIP(reg) potentials. The unregularised aPIP failed the vacancy test.

Model	Elastic constants (GPa)				Surface energy (J m^{-2})			Point defect (eV)
	B	C_{11}	C_{12}	C_{44}	(100)	(110)	(111)	vacancy
DFT	87.45	152.21	55.07	74.95	2.17	1.52	1.57	3.67
	Relative error (%)							
GAP	<1	−3	3	−7	−1	−1	−3	−2
aPIP(unreg)	8	5	12	−1	−6	−2	−5	—
aPIP(reg)	<1	3	−5	−4	−3	−3	−10	−4

relaxation of the vacancy using the aPIP(unreg) potential the optimiser failed to find a local minimiser. This is a manifestation of 'holes' in the fit which will be discussed in section 3.3. In comparison to SOAP-GAP the aPIP (reg) performs well in calculating the elastic constants and vacancy formation energy. The aPIP(reg) performs worse in the surface formation energies, most specifically the (111) direction, which we believe can only be rectified by using even higher body order terms Implementing this efficiently is a focus of future work.

3.3. Holes

As a first test of the 'transferability' of our regularised aPIP fits we determine whether the resulting potentials have any 'holes' in the sense of [31]: regions of unphysically low potential energy values. The importance of avoiding such behaviour is hard to overstate: in most molecular modelling applications, samples will be drawn using the *model* distribution, and due to the exponential amplification of the Boltzmann distribution at low and moderate temperature, holes can lead to catastrophic failure of models.

Having decomposed the total PES into low-dimensional components gives the option of searching for low-energy configurations in these individual components V_n. Specifically, we choose a minimal inner distance $r_0 < r'_{\text{cut}}$, i.e. below the inner cutoff. Then, for each n-body term V_n we compute an approximate minimum of $V_n(\mathbf{r}_1, ..., \mathbf{r}_n)$ over all clusters $(\mathbf{r}_1, ..., \mathbf{r}_n)$ with $r_0 \leqslant r_j \leqslant r_{\text{cut}}$, using Sobol sequences with a few million points. The results are summarized in table 7. It is interesting to note that the 'holes' in the unregularised Si fit are less severe. We speculate that this is due to the fact that the Si training set is much richer; see figure 9.

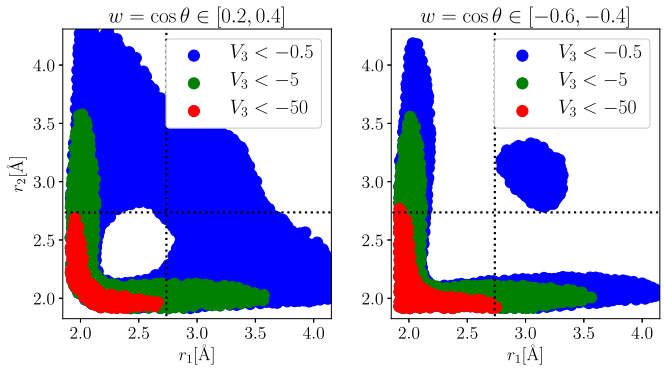


Figure 9. Holes in the 3-body potential V_3 for the unregularised W aPIP fit. Two slices through the potential energy landscape are plotted.

Table 7. 'Holes' in the unregularised aPIP (unreg) potentials.

	n	inf V_n aPIP(unreg)	aPIP(reg)
W	2	−273 eV	−0.08 eV
	3	−192 359 eV	−0.06 eV
	4	−4877 eV	−0.18 eV
Si	2	−0.66 eV	−0.08 eV
	3	−2934 eV	−0.6 eV
	4	−447 eV	−0.11 eV

In the unregularised fit we visualise the location of holes by plotting slices through the V_3 energy landscape in figure 9. This shows in particular that holes appear both in regions of large angles and moderate angles near the ground-state.

4. Generalisation

This section presents a series of tests to benchmark the generalisation performance of the aPIP and GAP models against a DFT reference. These tests were designed to probe configurational space far from the training data region and monitor each models' performance. We use the term 'generalisation' (also often called 'extrapolation' or 'transferability') in a loose sense and simply take it to mean 'evaluation far from the training set' which may technically also include *interpolation*—the distinction is tenuous in high dimension.

4.1. Tungsten

4.1.1. Interstitial

The tungsten aPIP(unreg) and aPIP(reg) models introduced in 3.2.1 were compared against the original SOAP-GAP model [47] and a DFT reference for the self-interstitial defect. The DFT parameters were chosen to be identical to those used in the original paper. These settings were: 600 eV cutoff energy, 0.03 $Å^{-1}$ kpoint spacing and 0.1 eV smearing width. The training set does not include interstitial data; see table 4. We formed the interstitial defect by inserting an atom in a DFT relaxed 54 atom tungsten bcc cell at the octahedral site $(\frac{1}{2}, 0, 0)$ of the primitive cell. The geometry was then relaxed using DFT. A linear path between the unrelaxed and relaxed configurations was created and the DFT, SOAP-GAP and aPIP models were evaluated along this path and are shown in figure 10.

This interstitial test probes small interatomic distances which are not contained in the training database and can therefore be strictly seen as an extrapolation test. Due to the smoothness prior, the SOAP-GAP model underestimates the energy difference along the interstitial path compared to the DFT reference. As expected, the unregularised aPIP model heavily oscillates along this test path since it explores configurations it was not fitted to. By contrast, the combination of integral regularisers, repulsive core and inner cut-offs result in a regularised aPIP model that shows an excellent match to the DFT curve. The level of agreement is likely fortuitous, but we expect that in general the correct repulsive nature of the potential is obtained due to having enforced this property in the two body function V_2.

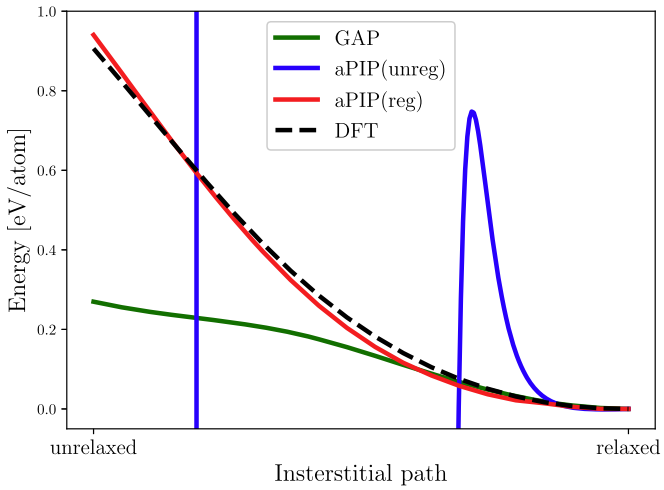


Figure 10. Linear path between unrelaxed and relaxed tungsten interstitial configuration. The unregularised aPIP fit explodes whereas the regularised aPIP closely matches the DFT reference.

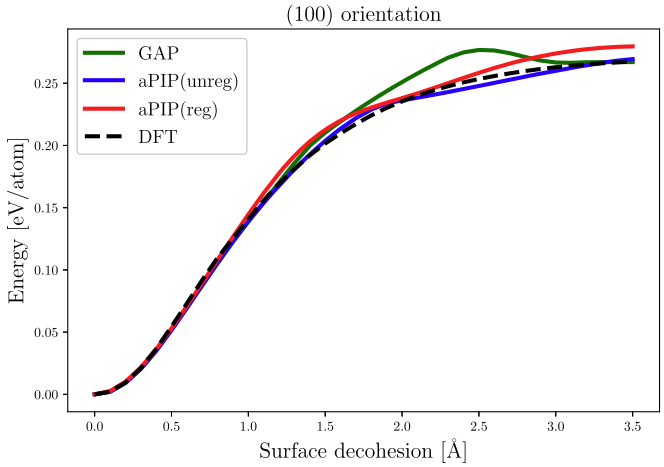


Figure 11. Decohesion curve in silicon bulk diamond along the (100) direction. The bulk diamond structure corresponds to the left end and the unrelaxed (100) surface is on the right.

4.2. Silicon

4.2.1. Surface decohesion

We set up a bulk Si diamond 10 × 1 × 1 supercell and increase the lattice vector length in the long direction while keeping the atomic positions fixed which in turn creates two surfaces [13]. Both the initial bulk structure and final surfaces are configurations that are well represented in the training set and fitted by the aPIP(unreg), aPIP(reg) and SOAP-GAP models to within 3 meV accuracy; see table 5. The configurations along the path in between are not contained in the database. Therefore this test can be seen as generalising in that it evaluates the potential on a path between two accurately fitted configurations.

Figure 11 shows that the end points, bulk diamond and (100) surface, are accurately fitted. However, the SOAP-GAP has a local maximum around 2.5 Å, unlike the DFT reference which shows a smooth and monotone transition along this path, a characteristic which both the aPIP(unreg) and aPIP(reg) mimic more accurately than SOAP-GAP.

4.2.2. Layer test

In this test we set up a bulk silicon diamond configuration and and gradually increase the interplanar spacing between the (111) layers of silicon. The configurations along this path are arguably unphysical but should, as the DFT model confirms, correspond to unstable high energy configurations. Past experience shows that poorly fitted potentials can have unphysically low energies for such configurations. The aPIP and SOAP-GAP models were evaluated on this path and are compared to the DFT reference in figure 12.

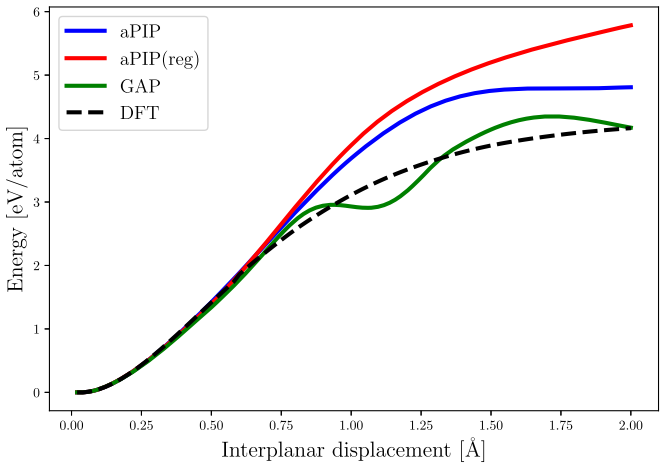


Figure 12. The layer test along which bulk silicon is split in layers of silicene. The SOAP-GAP model predicts a high energy local minima whereas the regularised aPIP model does not.

The SOAP-GAP model predicts a high energy local minimum along this path, which should be entirely unstable as shown by the DFT reference. The presence of such false high energy local minima are detrimental for applications such as random structure search [53] but also high temperature or high stress molecular dynamics. The aPIP(unreg) model has a much more shallow high energy local minimum, but the aPIP(reg) model show no minimum at all (while still not being quantitatively accurate in the unphysical region).

4.3. Titanium

Titanium is a difficult material to construct interatomic potentials for, due to its bonding chemistry being intermediate between covalent and metallic and we therefore chose it for our final test system. Here we want to explore how the different functional forms and regularisation strategies perform in the limit of *very little data*. The motivation for this is partly that the large size of the published data sets in the previous sections in and of itself acts like a regulariser, but also that in the future we wish to eliminate extensive sampling as a way to generate data sets. Rather, we would like to develop potential fitting frameworks that are explicitly regularised to the extent that a few judiciously chosen training configurations are sufficient to obtain good interatomic potentials.

We generated two very limited training sets, denoted by Set 1 and Set 2. To Set 1, we fit unregularised and regularised aPIP potentials denoted, respectively, by aPIP_1(unreg) and aPIP_1(reg), as well a SOAP-GAP potential denoted GAP_1. In addition we fit an unregularised aPIP potential, denoted aPIP_2 (unreg) to Set 2. The detailed potential and fitting parameters are described in the supplementary information.

Set 1 This set contains primitve cell bcc and hcp configurations, obtained by sampling the Boltzmann distribution with a temperature parameter set to 100 K as the lattice vectors (and in case of bcc, the relative positions of the two atoms) are varied. The obtained configurations were evaluated using CASTEP [48] with k-point spacing set to 0.015 Å^{-1}, 750 eV cutoff energy and 0.1 eV smearing width. In addition $3 \times 3 \times 3$ bcc and hcp supercells were added, and Phonopy [54] was used to generate the inequivalent finite displacements (one for each structure) of a single atom by a magnitude of 0.001 Å. These two configurations were evaluated with a larger k-point spacing equal to 0.03 Å^{-1}.

Set 2 This set contains Set 1 as well as additional finite displacement configurations analogous to those in Set 1, but now with a 0.01 Å displacement.

4.3.1. Cohesive energy

Figure 13 shows the energy versus volume curves for the titanium bcc and hcp phases for the various potentials all fitted to Set 1. The distribution of the atomistic volumes are on the lower panel and show that training data covers only volumes in the range between 15 and 17 Å^3/atom. In this region both aPIP models as well as the GAP model agree well with the DFT reference. For smaller and larger volumes the fitted potentials all deviate from the DFT reference as expected due to the lack of data. However, the unregularised aPIP shows a steep drop for small volumes and a second minimum for large volumes whereas the regularised aPIP and GAP mimic the DFT curve at least qualitatively, showing the benefits of regularisation.

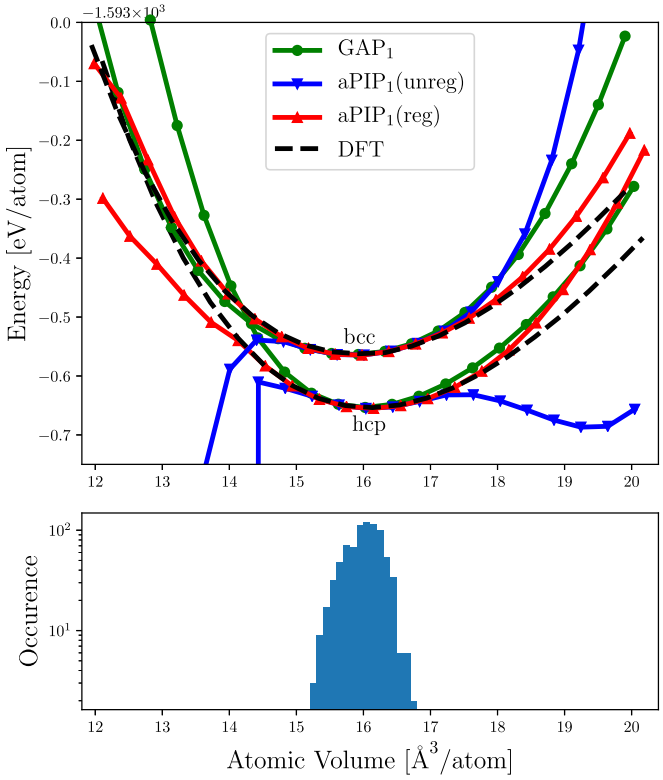


Figure 13. Energy versus volume plots for Titanium hcp and bcc phase. The volume histogram below shows the distribution of the training data.

4.3.2. Phonon spectrum

SOAP-GAP and aPIP models fitted to Set 1 and Set 2 were used to generate the phonon spectra for bcc and hcp titanium and are compared to the DFT reference in figure 14. Although $aPIP_1$(unreg) and $aPIP_1$(reg) were both fitted to Set 1, $aPIP_1$(unreg) failed catastrophically while $aPIP_1$(reg) produces a highly accurate phonon spectrum. The negative frequencies of the bcc phonon spectrum at the Γ points demonstrates the instability of the structure at 0 K.

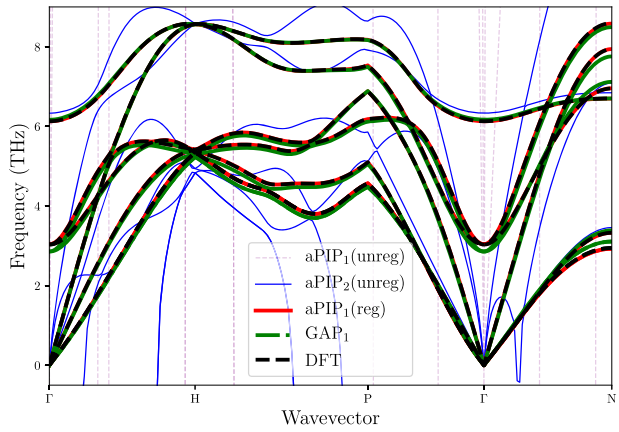
Of course an alternative, but much less controllable way to improve the regularity of a potential is to simply add more data to the training set. ML potentials in the literature, even if they are not explicitly regularised, can avoid the unphysical behaviour seen in here because they are fitted to large data sets. To see this effect, consider the potentials fitted to Set 2: $aPIP_2$(unreg) phonon spectra are somewhat improved, but they remain highly inaccurate quantitatively. We expect that much more training data would be required to accurately converge the phonon spectrum of an unregularised potential. We suggest instead that regularisation and a single displacement per crystal structure are enough to accurately reproduce phonon spectra.

4.3.3. Burgers' path

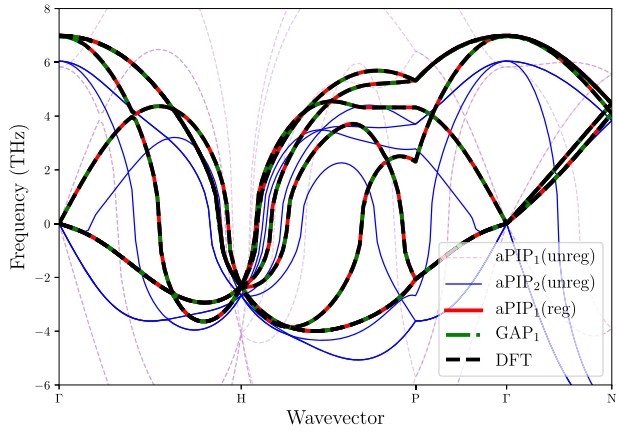
The Burgers' path is a pathway from the bcc to hcp crystal structure [55]. It consists of a shear deformation applied to the bcc structure followed by a shuffle of atomic layers resulting in a hcp structure [56].

We evaluated the Set 1 fits on the Burgers' path and the results are plotted against the DFT reference in figure 15. The energy per atom along the Burgers' path shows that the SOAP-GAP model overestimates the barrier by 30 meV along the transition from bcc to hcp. The aPIP models as well as the DFT reference do not show such a barrier. The training database contains bcc/hcp configurations sampled at a low temperature ($T = 100$ K) resulting in a database containing structures close to the relaxed bcc and hcp configurations. As expeced, figure 15 shows the aPIP and SOAP-GAP models both predict the energies for the hcp and bcc crystals accurately. However, the aPIP model manages to predict the energy along the middle part of the path more accurately compared to the DFT reference.

To analyse this we use body order expansion to investigate the Burgers' path test in a different way. By plotting the 3-body distances r_{12}, r_{13}, r_{23} of the primitive cell training database, Set 1, we can visualise the clustering of data in the V_3 space. Figure 16 shows the clustering of data around the relaxed bcc/hcp configurations and shows the Burgers' path connecting the two configurations as well. The training data was sampled at a low temperature ($T = 100$ K) and we therefore see a limited exploration of data away from the



(a) Titanium hcp phonon spectrum



(b) Titanium bcc phonon spectrum

Figure 14. Phonon spectra computed using the aPIP, SOAP-GAP models and DFT reference. Regularisation increases the accuracy of the aPIP models, similarly to the addition of FDC data.

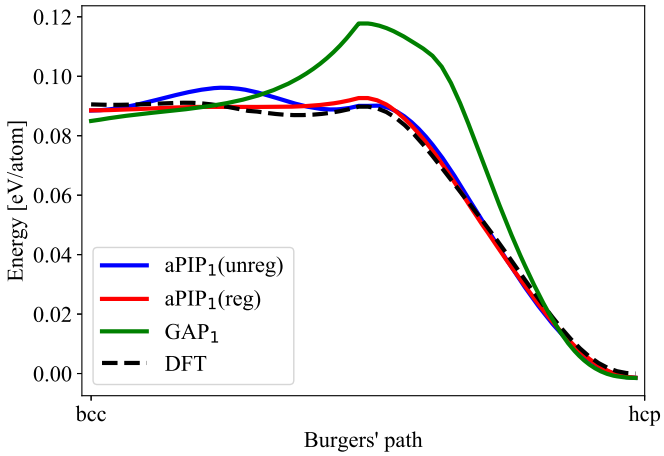


Figure 15. Burgers' path comparing aPIP, SOAP-GAP and DFT reference. The SOAP-GAP model correctly predicts the relaxed bcc/hcp energies at both ends of the path but overestimates the energy along the remainder of the path.

relaxed configurations. The Burgers' path clearly explores configurations in this V_3 space where there is little to no data present, showing that even in this low dimension this test can be considered as a non-trivial generalisation.

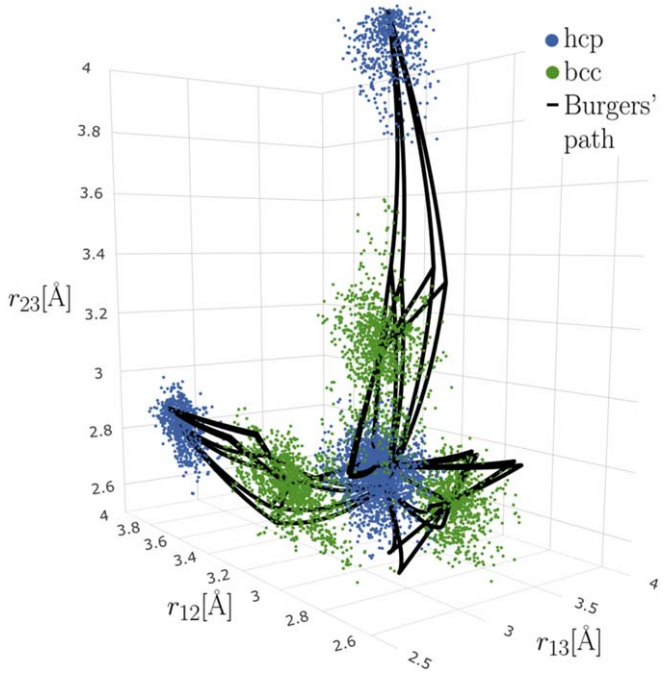


Figure 16. Visualisation of the sparsity of the Ti training sets: each datapoint represents a 3-body cluster, described through interatomic distances r_{12}, r_{13}, r_{23}, contained in the primitive cell configurations of Set 1. The Burgers' path explores regions of 3-body space far from the regions with datapoints, which demonstrates that the Burgers' path test requires significant generalisation away from the training set.

We believe that the lack of accuracy along the Burgers' path of SOAP-GAP model is due to the lack of training data in that region. It is a manifestation of high dimensional fits giving answers that are still regular, but have uncontrolled errors when extrapolating away from the training database. The aPIP models perform significantly better in this test because, as we propose, models fitted in low dimension (e.g. the dimensionality of the body orders) will in general perform better in generalisation tests compared to high dimensional fits such as SOAP-GAP models.

It will also be interesting to systematically explore the relative benefits and possible unification of regularisation and active-learning style techniques that bring in new data in previously unexplored regions [31, 57].

5. Conclusion

The purpose of this paper was two-fold: firstly, we developed *atomic PIPs* (aPIPs), a generalisation of PIPs [40], interatomic potentials constructed from permutation invariant polynomials for material systems by applying the PIP construction to atomic body ordered expansions of the total energy and endowing them with usual cutoff mechanisms. By fitting the polynomial coefficients to solid training sets (rather than clusters in vacuum) we were able to obtain an accuracy comparable with the SOAP-GAP models for tungsten [47] and silicon [13] (on a non-trivial subset of the full training set) using low body-orders, four or five, which are still at least an order of magnitude faster to evaluate than SOAP-GAP.

Secondly, we studied the generalisation (extrapolation) properties of the aPIPs. We developed novel regularisation mechanisms that exploit the low-dimensional structure of the body-ordered terms to ensure correct qualitative behaviour of the potentials, such as smoothness, well away from the training set. We showed that such a regularisation is crucial to achieve the extrapolation properties of the Gaussian process based SOAP-GAP [13, 47]. Indeed, our regularisation techniques are amenable to fine-tuning which enabled us to significantly improve on the SOAP-GAP model in several tests. Thus, we have established that our framework provides a novel 'tool kit' for fitting interatomic potentials for materials with high accuracy and excellent transferability, across a wide range of bonding chemistries.

The silicon, tungsten and titanium datasets are openly available at http://libatoms.org/Home/DataRepository, the aPIP framework can be found at https://github.com/cortner/NBodyIPs.jl.

Acknowledgments

GC and C vd O acknowledge the support of UKCP grant number EP/K014560/1. C vd O would like to acknowledge the support of EPSRC (Project Reference: 1971218) and Dassault Systèmes UK. CO and GD are supported by Leverhulme Research Project Grant RPG-2017-191.

ORCID iDs

Cas van der Oord https://orcid.org/0000-0003-1845-0387
Geneviève Dusson https://orcid.org/0000-0002-7160-6064
Gábor Csányi https://orcid.org/000-0002-8180-2034
Christoph Ortner https://orcid.org/0000-0003-1498-8120

References

[1] Finnis M 2004 *Prog. Mater Sci.* **49** 1
[2] Parr R G 1980 *Horizons of Quantum* ed K Fukui and B Pullman (Dordrecht: Springer) pp 5–15
[3] Ercolessi F and Adams J B 1994 *Europhys. Lett.* **26** 583
[4] Murrell J N 1984 *Molecular Potential Energy Functions* (New York: Wiley)
[5] Atkins P and Friedman R 2011 *Molecular Quantum Mechanics* (Oxford: OUP)
[6] Huang X, Braams B J and Bowman J M 2005 *J. Chem. Phys.* **122** 044308
[7] Jain A K, Mao J and Mohiuddin K M 1996 *Computer* **29** 31
[8] Bishop C M 2006 *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Berlin: Springer)
[9] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
[10] Rasmussen C E and Williams C K I 2005 *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (Cambridge, MA: MIT Press)
[11] Bartók A P and Csányi G 2015 *Int. J. Quantum Chem.* **115** 1051
[12] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[13] Bartók A P, Kermode J, Bernstein N and Csányi G 2018 *Phys. Rev.* X **8** 041048
[14] Partridge H and Schwenke D W 1997 *J. Chem. Phys.* **106** 4618
[15] Bramley M J and Carrington T Jr 1993 *J. Chem. Phys.* **99** 8519
[16] Cencek W, Szalewicz K, Leforestier C, van Harrevelt R and van der Avoird A 2008 *Phys. Chem. Chem. Phys.* **10** 4716
[17] Xie Z, Braams B J and Bowman J M 2005 *J. Chem. Phys.* **122** 224307
[18] Babin V, Leforestier C and Paesani F 2013 *J. Chem. Theory Comput.* **9** 5395–403
[19] Qu C, Yu Q and Bowman J M 2018 *Annu. Rev. Phys. Chem.* **69** 151
[20] Griebel M 2006 *Foundations of Computational Mathematics (FoCM05)* ed L Pardo *et al* (Cambridge: Cambridge University Press) pp 106–61
[21] Rabitz H and Aliş Ö F 1999 *J. Math. Chem.* **25** 197
[22] Stillinger F H and Weber T A 1985 *Phys. Rev.* B **31** 5262
[23] Nguyen T T, Székely E, Imbalzano G, Behler J, Csányi G, Ceriotti M, Götz A W and Paesani F 2018 *J. Chem. Phys.* **148** 241725
[24] Glielmo A, Zeni C and Vita A De 2018 *Phys. Rev.* B **97** 184307
[25] Sharma A R, Wu J, Braams B J, Carter S, Schneider R, Shepler B and Bowman J M 2006 *J. Chem. Phys.* **125** 224306
[26] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[27] Huo H and Rupp M 2017 arXiv:1704.06439
[28] Behler J 2017 *Angew. Chem. Int. Ed.* **56** 12828
[29] Behler J 2011 *J. Chem. Phys.* **134** 074106
[30] Deringer V L and Csányi G 2017 *Phys. Rev.* B **95** 094203
[31] Nandi A, Qu C and Bowman J M 2019 *J. Chem. Theory Comput.* **15** 2826
[32] Shapeev A V 2016 *Multiscale Model. Simul.* **14** 1153
[33] Drautz R 2019 *Phys. Rev.* B **99** 014104
[34] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[35] Thompson A, Swiler L, Trott C, Foiles S and Tucker G 2015 *J. Comput. Phys.* **285** 316
[36] Moriarty J A 1977 *Phys. Rev.* B **16** 2537
[37] Moriarty J A 1982 *Phys. Rev.* B **26** 1754
[38] Moriarty J A 1988 *Phys. Rev.* B **38** 3199
[39] Nelson L J, Ozoliņš V, Reese C S, Zhou F and Hart G L 2013 *Phys. Rev.* B **88** 155105
[40] Braams B J and Bowman J M 2009 *Int. Rev. Phys. Chem.* **28** 577
[41] Qu C and Bowman J M 2019 *J. Chem. Phys.* **150** 141101
[42] Shapeev A 2016 *Multiscale Model. Simul.* **14** 1153
[43] Derksen H and Kemper G 2015 *Computational Invariant Theory* (Berlin: Springer)
[44] Qu C, Prosmiti R and Bowman J M 2013 *Theor. Chem. Acc.* **132** 1413
[45] Schmelzer A and Murrell J N 1985 *Int. J. Quantum Chem.* **28** 287
[46] Bosma W, Cannon J and Playoust C 1997 *J. Symb. Comput.* **24** 235
[47] Szlachta W J, Bartók A P and Csányi G 2014 *Phys. Rev.* B **90** 104108
[48] Clark S J, Segall M D, Pickard C J, Hasnip P J, Probert M I J, Refson K and Payne M C 2005 *Z. Kristallogr.* **220** 567–70
[49] Dragoni D, Daff T D, Csányi G and Marzari N 2018 *Phys. Rev. Mater.* **2** 1939
[50] Chan T F 1987 *Linear Algebra Appl.* **88–89** 67
[51] Niederreiter H 1988 *J. Number Theory* **30** 51–70
[52] Ziegler J F, Biersack J P, Littmark U *et al* 1985 *The Stopping and Range of Ions in Solids* vol 1 (New York: Pergamon)

[53] Pickard C J and Needs R J 2011 *J. Phys.: Condens. Matter* **23** 053201
[54] Togo A and Tanaka I 2015 *Scr. Mater.* **108** 1
[55] Burgers W 1934 *Physica* **1** 561
[56] Caspersen K J, Lew A, Ortiz M and Carter E A 2004 *Phys. Rev. Lett.* **93** 115501
[57] Podryabinkin E V and Shapeev A V 2017 *Comput. Mater. Sci.* **140** 171